The infinitely many genes model with horizontal gene transfer

Franz Baumdicker[1] and Peter Pfaffelhuber[1,2]

January 29, 2013

Abstract

The genome of bacterial species is much more flexible than that of eukaryotes. Moreover, the distributed genome hypothesis for bacteria states that the total number of genes present in a bacterial population is greater than the genome of every single individual. The *pangenome*, i.e. the set of all genes of a bacterial species (or a sample), comprises the core genes which are present in all living individuals, and accessory genes, which are carried only by some individuals. In order to use accessory genes for adaptation to environmental forces, genes can be transferred horizontally between individuals. Here, we extend the *infinitely many genes model* from Baumdicker, Hess and Pfaffelhuber (2010) for horizontal gene transfer. We take a genealogical view and give a construction – called the *Ancestral Gene Transfer Graph* – of the joint genealogy of all genes in the pangenome. As application, we compute moments of several statistics (e.g. the number of differences between two individuals and the gene frequency spectrum) under the infinitely many genes model with horizontal gene transfer.

Keywords and phrases: Prokaryote, bacterial evolution, coalescent, gene frequency spectrum, pangenome

AMS Subject Classification: 92D15, 60J70, 92D20 (Primary); 60K35 (Secondary)

1 Introduction

Today, many prokaryotic species (i.e. bacteria and archea) are known to have highly flexible genomes (e.g. Tettelin et al., 2005; Ehrlich et al., 2005; Tettelin et al., 2008; Koonin and Wolf, 2012). Unlike in eukaryotes, genes can be transferred horizontally (i.e. without a direct relationship between donor and recipient) between prokaryotic individuals of either different or the same population. As a result, gene content can differ substantially between strains from the same population. For example, the pathogenic strain *E. coli* O157:H7 carries 1387 genes which are absent in the commensal strain *E. coli* K-12 (Perna et al., 2001). This huge variation in gene content led to the concepts of the *distributed*

[1]Abteilung für Mathematische Stochastik, Albert-Ludwigs University of Freiburg, Eckerstr. 1, D–79104 Freiburg, Germany
[2]e-mail: p.p@stochastik.uni-freiburg.de

genome of bacteria and their *pangenome* (Tettelin et al., 2005; Ehrlich et al., 2005). In datasets, genes present in all genomes of a taxon are called *core genes* while genes present in only some but not all individuals comprise the *accessory genome*. The latter set of genes is further split into the medium-frequency *shell* of genes and the *cloud* of genes of low frequency Koonin and Wolf (2008).

In order to understand the growing amount of genomic data from bacterial species, classical population genetic theory – using mutation, selection, recombination and genetic drift as main evolutionary forces – must be extended in order to include realistic mechanisms of horizontal gene transfer (HGT). Since genomic data from prokaryotic species has become abundant only recently, HGT can in particular be seen as a newly discovered evolutionary factor (Doolittle, 1999; Koonin et al., 1997). Current estimates for the amount of genes which are horizontally transferred are at least as high as 32% (Koonin et al., 2001; Dagan and Martin, 2007). It may even be argued that this number is still a lower bound because only a fraction of all events of HGT can be seen in data, either because the transferred gene is subsequently lost or the pattern is in accordance to vertical gene transfer (Gogarten et al., 2002).

The *tree of life* has become a classical way of thinking about inheritance since Darwin's *Origin of Species*. However, the abundance of HGT counteracts the tree-like structures evolutionary biologists like to think about. Results are phylogenetic networks, which display at the same time the joint evolutionary fate of many genes (Huson and Scornavacca, 2011; Dagan, 2011), in addition to other reticulate events such as hybridization and incomplete lineage sorting. It is becoming clear that any genealogical tree of bacteria which have a flexible genome is at most a tree of 1% of all genetic material (Dagan and Martin, 2006), which may eventually lead to a paradigm shift in evolutionary biology of prokaryotes (Koonin and Wolf, 2012).

Unraveling the amount of HGT shaping bacterial diversity can today be tackled using a growing amount of genomic data. In particular, several datasets from closely related strains, which are of the same bacterial species are available today (Medini et al., 2005; Tettelin et al., 2005, 2008). In addition, many different approaches have led to a number of methods to estimate HGT rates and identify the corresponding genes (Lawrence and Ochman, 2002; Kunin and Ouzounis, 2003; Nakhleh et al., 2005; Linz et al., 2007; Didelot et al., 2010). However, theoretical work on the population genomics of HGT is still in its infancy. In order to include HGT in population genetic models, some scenarios have to be distinguished according to the following basal mechanisms: (a) Transformation, which is the uptake of genetic material from the environment. (b) Transduction, which describes the infection of a prokaryote by a lysogenic virus (phage) which provides additional genetic material that can be built in the bacterial genome. (c) Conjugation, which is also termed *bacterial sex*, which requires a direct link (pilus) between two bacterial cells and leads to exchange of genetic material. In addition, small virus-like elements called Gene Transfer Agents (GTAs) have been found which may become even more important for the amount of horizontal genetic exchange in some species (McDaniel et al., 2010). Another mechamism of horizontal gene transfer are due to mobile genetic elements like plasmids, gene cassettes and transposons, which transfer genes even within a single individual (de la Cruz and Davies, 2000). Most importantly, when considering horizonal gene transfer by (a), (b) and (c), transformation

(and to a lesser extend also transduction) transfers genes mainly from between distantly related species, while conjugation only works for bacteria from closely related species.

A first approach of the population genomics of bacteria was made by Novozhilov et al. (2005), extending a model from Berg and Kurland (2002). Here, a birth and death process is used in order to describe the evolution of the frequency of a single gene under selection under within-population horizontal transmission ("infection"), mutation (leading to loss of the gene) and population size changes. However, this study is limited since only a single gene is considered, but bacterial genomes are comprised of several hundreds of genes, each of which may be under selection and horizontal gene transfer. In Mozhayskiy and Tagkopoulos (2012), a simulation study was carried out, taking selective forces into account which arise from gene regulatory networks, i.e. epistasis of presence and absence of genes. Finally, Vogan and Higgs (2011) present a macro-evolutionary model and conclude that HGT was probably favorable in early evolution since loss of genes is frequent, but later, when genomes are rather optimized, HGT is not favorable and gene losses are rarer.

In Baumdicker et al. (2010) (see also Baumdicker et al., 2012), we presented the *infinitely many genes model*, a population genomic model which includes HGT from different species, e.g. by transformation, but no HGT within species. It accounts for *gene gain* (or gene uptake) from the environment (at rate $\theta/2$) along the genealogical tree which describes the relationships between the individuals of the population. The term *gene gain* covers HGT from other species as well as gene genesis, because from the perspective of a species under consideration these two mechanisms are indistinguishable. Pseudogenization may lead to deletion of genes and is incorporated by *gene loss* (at rate $\rho/2$). The model uses the coalescent (Kingman, 1982; Hudson, 1983) as underlying genealogy instead of a fixed (phylogenetic) tree. On the latter, the same two mechanisms were studied already by Huson and Steel (2004). A variant of the infinitely many genes model was introduced by Haegeman and Weitz (2012), who couple gene gain and loss events in order to obtain a genome of constant size. However, this is in contrast to available data, since flexible genomes of bacteria usually come with different genome sizes of up to 50%. An interesting extension of the infinitely many genes model was studied in Collins and Higgs (2012). Here, different random trees were used as underlying genealogies as well as different classes of genes, each class with its own rate of gene gain and loss. It was found that the coalescent produces a good fit with data, and it is likely that the rate of gene gain and loss depends on the gene.

In the present paper, we extend the infinitely many genes model in order to incorporate events of intraspecies horizontal gene transfer. We stress that HGT in bacteria differs from crossover recombination in eukaryotes, since only single, non-homologous, genes are horizontally transferred in bacteria, while only homologous genomic regions are transferred by recombination in eukaryotes. Accordingly, since we aim at a genealogical picture of HGT in bacteria, the ancestral recombination graph (Hudson, 1983; Griffiths and Marjoram, 1997) as an extension of the coalescent, cannot be used. Rather, we model HGT such that each gene present in the population comes with its own events of HGT, resulting in the Ancestral Gene Transfer Graph (AGTG). In the limit of large population sizes, we compute moments of several quantities of interest. The gene

frequency spectrum – see Theorem 1 – describes the amount of genes present in k out of n individuals. In Theorem 2, we give our results for the expectations of the average number of genes per individual, and the average number of pairwise differences and the total number of genes, respectively. Calculations which give the variances of some of these quantities, can be carried out using the AGTG and are given in Theorem 3.

The paper is organized as follows: In Section 2, we introduce the infinitely many genes model with horizontal gene transfer. After stating our results in Section 3, we introduce the AGTG in Section 4. Proofs of the Theorems are then given in Section 5.

2 The model

We introduce two different views on the same model. In this section, we describe a Moran model forwards in time, including events of gene gain, gene loss and horizontal transfer of genes. Later, in Section 4 we describe how to obtain the distribution of genes in equilibrium using a genealogy-based approach.

We consider the following model for bacterial evolution: Each bacterial cell carries a set of *genes* and every gene belongs either to the *core genome* or to the *accessory genome*. The infinite set $I := [0, 1]$ is the set of conceivable accessory genes and $\mathcal{G}_c$ with $\mathcal{G}_c \cap I = \emptyset$ is the core genome. A population of constant size consists of N individuals (bacterial cells). We model the accessory genome of individual i at time t by a finite counting measure $\mathcal{G}_i^N(t)$ on I. We will identify finite counting measures with the set of atoms, i.e. we write $u \in \mathcal{G}_i^N(t)$ if $\langle \mathcal{G}_i^N(t), 1_u \rangle \geq 1$. The dynamics of the model is such that $\langle \mathcal{G}_i^N(t), 1_u \rangle \leq 1$ for all i and $u \in [0, 1]$, almost surely.

The population evolves according to Moran dynamics. That is, time is continuous and every (ordered) pair of individuals undergoes a resampling event at rate 1. Here, in each resampling event between individuals i and j, one bacterium is chosen at random, produces one offspring which replaces individual j (such that the population size stays constant). The offspring carries the same genes as the parent, i.e. if an offspring of i replaces j at time t, we have $\mathcal{G}_j^N(t) = \mathcal{G}_i^N(t-)$. In addition to such resampling events, the following (independent) events occur:

1. *Gene loss*: For gene $u \in \mathcal{G}_i^N(t-)$ in individual i, at rate $\rho/2$, we have $\mathcal{G}_i^N(t) = \mathcal{G}_i^N(t-) \setminus \{u\}$, i.e. gene u is lost from $\mathcal{G}_i^N(t)$.

2. *Gene gain*: For every individual i, at rate $\theta/2$, choose U uniform in $[0, 1]$ and set $\mathcal{G}_i^N(t) = \mathcal{G}_i^N(t-) \cup \{U\}$, i.e. every individual gains an (almost surely) new gene at rate $\theta/2$.

3. *Horizontal gene transfer*: For every (ordered) pair of individuals (i, j) and $u \in \mathcal{G}_i^N(t)$, a horizontal gene transfer event occurs at rate $\gamma/2N$. For such an event, set $\mathcal{G}_j^N(t) = \mathcal{G}_j^N(t-) \cup \{u\}$, i.e. individual i is the *donor* of gene u (the *transferred gene*) to the *acceptor* j.

Horizontal gene transfer events can as well be written in the measure-valued notation as $\mathcal{G}_j^N(t) = (\mathcal{G}_j^N(t-) + \delta_u) \wedge 1)$. The '$\wedge 1$'-term indicates that we do

not model paralogous genes, i.e. horizontal gene transfer events have no effect if the acceptor individual j already carries the transferred gene.

Definition 2.1 (Moran model with horizontal gene transfer)**.** *We refer to* $(\mathcal{G}_1^N(t), \dots, \mathcal{G}_N^N(t))_{t\geq 0}$ *undergoing the above dynamics as the* Moran model for bacterial genomes with horizontal gene flow.

Remark 2.2 (Equilibrium)**.** It can be shown that the Moran model of size N for bacterial genomes with horizontal gene flow is Harris recurrent and hence, has a unique equilibrium. We denote the corresponding limit by $\mathcal{G}_1^N := \mathcal{G}_1^N(\infty), ..., \mathcal{G}_N^N := \mathcal{G}_N^N(\infty)$.

Remark 2.3 (Extensions of the model)**.** Although the above introduced model for bacterial evolution with horizontal gene flow is among the most realistic models for bacterial evolution available so far, several extensions are well possible, e.g.:

- *Gain, loss and transfer of multiple genes*: The exact mechanisms of gene gain, loss and HGT are still under study. However, it seems clear that several genes can be gained or lost at once.
- *Gene families*: Frequently, a single gene is present not only once but several times in a bacterial genome. The reason can either be a copying event along its ancestral line, or the gene is introduced by HGT although it was already present.
- *Gene synteny*: The order of genes in the genome is called gene synteny. In our model, the synteny of genes is not modeled, but can be observed in genomic data. Above all, gene synteny can give hints of events of horizontal gene transfer, since the order of genes can be different in donor and acceptor.
- *Mobile genetic elements*: There are parts of the genome like mobile elements which are more likely to be transferred horizontally. Examples are transposons, plasmids and gene cassettes, i.e. horizonral gene transfer even within a single cell can be considered.

Remark 2.4 (Measure-valued version)**.** In mathematical population genetics, one frequently studies models of finite populations forward in time, constructs their diffusion limit – most often a Fleming–Viot measure-valued diffusion – and only afterwards uses genealogical relationships in order to compute specific properties of the underlying forwards model. We take another route here and leave the construction of the infinite model forwards in time for future research. Here, one would have to consider the set of counting measures on $[0, 1]$ as a type space (which is locally compact), and define the current state of a finite population as the empirical measure of types on this state space. Constructing the diffusion limit then gives the measure-valued diffusion. Since our interest lies in seeing the effects of horizontal gene transfer on summary statistics (see Theorems 1–3), we do not follow this route within the current manuscript.

We are mainly interested in large populations. The corresponding limit is usually referred to as large population limit in the population genetic literature. The following result of the Moran model with HGT will already be useful in various applications.

Lemma 2.5 (The frequency path of a single gene)**.** *Let $X^N(t)$ be the frequency of gene u at time t in the Moran model for bacterial genomes with horizontal gene flow of size N with $X^N(0)$ such that $X^N(0) \xrightarrow{N\to\infty} x$. Then, in the large population limit, $N \to \infty$, the process $(X^N(t))_{t\geq 0}$ converges weakly to the solution of the SDE*

$$dX = \big(-\tfrac{\rho}{2}X + \tfrac{\gamma}{2}X(1-X)\big)dt + \sqrt{X(1-X)}dW \tag{2.1}$$

with $X(0) = x$ for some Brownian motion W.

Remark 2.6 (The diffusion (2.1) in population genetics)**.** The diffusion (2.1) also appears in population genetics models including selection (see e.g. Kimura, 1964; Ewens, 2004; Durrett, 2008). In the present setting, the term proportional to $X(1-X)dt$ appears because horizontal gene flow increases the frequency of the gene by a rate which is proportional to the number of possible donor/acceptor-pairs of individuals.

Due to the close connection of horizontal gene transfer with selective models, a comparison to recent work is appropriate. In particular, the theory for the frequency spectrum in selective models with irreversible mutations is carried out in Fisher (1930); Wright (1938); Kimura (1964, 1969). We rederive these results in our proof of Theorem 1 below, but we stress that the genealogical interpretation we give is derived with a special focus on horizontal gene flow, but not to the selective case.

Proof of Lemma 2.5. First note that gene loss reduces X^N with rate proportional to NX^N and $\rho/2$. Second, horizontal gene transfer increases X^N with rate proportional to $\gamma/(2N)$ and to the number of pairs where the horizontal gene transfer events has an effect, $N^2X^N(1-X^N)$. By construction, the evolution of frequencies of gene u is a Markov process with generator

$$\begin{aligned}(G^N f)(x) &= N(N-1)x(1-x)\Big(\frac{1}{2}f(x+1/N) + \frac{1}{2}f(x-1/N) - f(x)\Big) \\ &\quad - \frac{\rho N x}{2}(f(x-1/N) - f(x)) + \frac{\gamma N^2 x(1-x)}{2N}(f(x+1/N) - f(x)) \\ &\xrightarrow{N\to\infty} \frac{1}{2}x(1-x)f''(x) + (-\frac{\rho}{2}x + \gamma x(1-x))f'(x)\end{aligned}$$

for $f \in \mathcal{C}^2([0,1])$. Using e.g. standard results from Ewens (2004) it is now easy to show (weak) convergence to the diffusion (2.1). □

3 Results on Summary Statistics

Consider a sample $\mathcal{G}_1^N, \ldots, \mathcal{G}_n^N$ of size n taken from the Moran model of size N in equilibrium. We introduce several statistics under the above dynamics:

- The *average number of genes (in the accessory genome)* is given by

$$A^{(n)} := A^{(n,N)} := \frac{1}{n}\sum_{i=1}^{n} |\mathcal{G}_i^N| \tag{3.1}$$

  where $|\mathcal{G}_i^N| := \langle \mathcal{G}_i^N, 1\rangle$ is the total number of accessory genes in individual i.

- The *average number of pairwise differences* is given by

$$D^{(n)} := D^{(n,N)} := \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} |\mathcal{G}_i^N \setminus \mathcal{G}_j^N| \tag{3.2}$$

  where $\mathcal{G}_i^N \setminus \mathcal{G}_j^N := (\mathcal{G}_i^N - \mathcal{G}_j^N)^+$ are the genes present in i but not in j.

- The *size of the accessory genome* is given by

$$G^{(n)} := G^{(n,N)} := \Big| \bigcup_{i=1}^{n} \mathcal{G}_i^N \Big| \tag{3.3}$$

  where $\bigcup_{i=1}^n \mathcal{G}_i^N = \Big(\sum_{i=1}^n \mathcal{G}_i^N \Big) \wedge 1$ is the set of genes present in any individual from the sample.

- The *gene frequency spectrum (of the accessory genome)* is given by $G_1^{(n)} := G_1^{(n,N)}, \ldots, G_n^{(n)} := G_n^{(n,N)}$, where

$$G_k^{(n)} := G_k^{(n,N)} := |\{u \in I : u \in \mathcal{G}_i^N \text{ for exactly } k \text{ different } i\}|. \tag{3.4}$$

Remark 3.1 (Notation)**.** In the following results, we will suppress the superscript N of the population size. Instead, we query that our results hold in the *large population limit*. E.g. if we say that (3.6) holds in the large population limit, we really mean that

$$\mathbb{E}[A^{(n,N)}] \xrightarrow{N \to \infty} \frac{\theta}{\rho} \left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(1+\rho)_m} \right).$$

The proofs of all results presented here are given in Section 5. For first moments, we provide proofs using diffusion theory and Lemma 2.5. For second moments, we rely on the Ancestral Gene Transfer Graph (AGTG) of Section 4. Since the proofs of the results are either using Lemma 2.5 or the AGTG or both, we formulate the following three Theorems.

Theorem 1 (Gene frequency spectrum)**.** *Consider a sample of size n taken from the Moran model for bacterial genomes with horizontal gene flow with $\rho > 0, \theta > 0, \gamma \ge 0$ in equilibrium. Then, in the large population limit, it holds that*

$$\mathbb{E}[G_k^{(n)}] = \frac{\theta}{k} \frac{(n)_{\underline{k}}}{(n-1+\rho)_{\underline{k}}} \Big(1 + \sum_{m=1}^{\infty} \frac{(k)_{\bar{m}} \gamma^m}{(n+\rho)_{\bar{m}} m!} \Big) \tag{3.5}$$

with $(a)_{\bar{b}} := a(a+1)\cdots(a+b-1)$ and $(a)_{\underline{b}} := a(a-1)\cdots(a-b+1)$.

Theorem 2 (More sample statistics)**.** *Under the same assumptions as in Theorem 1,*

$$\mathbb{E}[A^{(n)}] = \frac{\theta}{\rho} \left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(1+\rho)_{\bar{m}}} \right), \tag{3.6}$$

$$\mathbb{E}[D^{(n)}] = \frac{\theta}{1+\rho} \left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(2+\rho)_{\bar{m}}} \right), \tag{3.7}$$

$$\mathbb{E}[G^{(n)}] = \theta \sum_{k=0}^{n-1} \frac{1}{k+\rho} + \theta \sum_{m=1}^{\infty} \frac{\gamma^m}{m} \Big(\frac{1}{(\rho)_{\bar{m}}} - \frac{1}{(n+\rho)_{\bar{m}}} \Big) \tag{3.8}$$

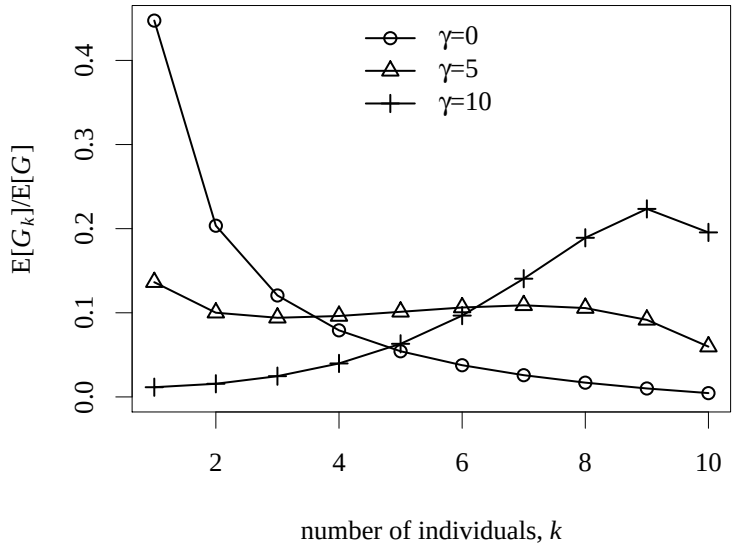

Figure 1: The expected gene frequency spectrum from Theorem 1 is highly dependent of γ, the rate of horizontal gene flow. For high values of γ, most genes are in high frequency, leading to a closed pangenome. We use $\rho = 2$ in the figure.

in the large population limit.

Theorem 3 (Second moment of the number of genes)**.** *Under the same assumptions and in the large population limit as in Theorem 1, we have, in the limit $\gamma \to 0$,*

$$\mathbb{V}[A^{(1)}] = \frac{\theta}{\rho}\Big(1 + \frac{1}{1+\rho}\gamma + \Big(\frac{1}{(1+\rho)(2+\rho)} + \frac{\theta}{(1+\rho)^2(3+2\rho)(2+7\rho+6\rho^2)}\Big)\gamma^2\Big) + \mathcal{O}(\gamma^3), \tag{3.9}$$

$$\mathbb{V}[D^{(2)}] = \frac{\theta}{1+\rho}\Big(\frac{1}{2} + \frac{\theta}{(1+\rho)(1+2\rho)} + \Big(\frac{1}{2(2+\rho)} + \theta\frac{2(12+110\rho+248\rho^2+209\rho^3+60\rho^4)}{(1+\rho)(2+\rho)(1+2\rho)^2(3+2\rho)(2+3\rho)(6+5\rho)}\Big)\gamma\Big) + \mathcal{O}(\gamma^2). \tag{3.10}$$

Remark 3.2 (Open and closed pangenomes)**.** Recently, the concepts of *open* and *closed* pangenomes were introduced (Medini et al., 2005). If, after sequencing a finite number of genomes, all genes present in the population are found, one speaks of a closed pangenome. If new genes are found even after sequencing many cells, the pangenome is *open*. It is not hard to see that high values of γ imply that most genes are in high-frequency. In other words, sequencing a new individual hardly leads to new genes which were not seen before. This impact of openness and closedness of the pangenome can as well be seen from Figure 1.

4 The Ancestral Gene Transfer Graph

Since the seminal work of Kingman (1982) and Hudson (1983), the genealogical view is a powerful tool in the analysis of population genetic models. Here, we will give a genealogical construction in order to obtain the distribution of $\mathcal{G}_1^N, ...\mathcal{G}_n^N$ for a sample of size $n \in \mathbb{N}$ in the large population limit of the Moran model with horizontal gene transfer in equilibrium. The resulting genealogy is denoted the *Ancestral Gene Transfer Graph* (AGTG). In this random graph, every ancestral line splits at constant rate $\gamma/2$ per gene due to a *potential* gene transfer event. We note that such events leading to potential ancestors are well-known for the ancestral selection graph (ASG) of Neuhauser and Krone (1997) and Krone and Neuhauser (1997). However, potential ancestors in the ASG arise by fitness differences within the population while the potential ancestors within the AGTG may take effect by events of horizontal gene transfer.

We start with the construction of the genealogy for a single gene and come to the full picture including all genes afterwards.

Definition 4.1 (The AGTG for a single gene)**.** *Consider a random graph $\mathcal{A}_n$ which arises as follows: Starting with n lines, denoted $i = 1, ..., n$,*

- *each (unordered) pair of lines coalesces at rate 1,*
- *each line disappears at rate $\rho/2$ (meaning that the gene was lost),*
- *each line splits in two lines at rate $\gamma/2$ (meaning that the gene was horizontally transferred from another individual, such that the gene can now have two different origins).*

Sample a single point E uniformly at random according to the length measure from the graph. (This point determines the time when the gene under consideration was gained.) For every line $1, ..., n$, let $G_i = 1$ if there is a direct (i.e. increasing in time) path from i to E and $G_i = 0$ otherwise. Then, $(G_1, ..., G_n)$ is denoted the gene distribution of a single gene read off from the AGTG.

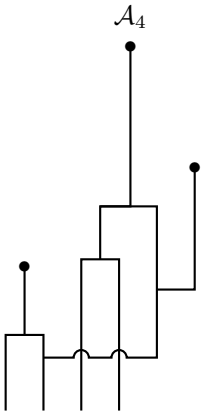

Figure 2: In the construction of the AGTG for a single gene $\mathcal{A}_4$, start with 4 lines at the bottom of the figure. Every pair of lines coalesces with rate 1, and every line splits at rate $\gamma/2$ and disappears (marked by $\bullet$) at rate $\rho/2$.

For later use, we show that all moments of the length of the AGTG are finite. In particular, the length is almost surely finite, and the uniform distribution according to the length measure, from which E is picked, is well-defined.

Lemma 4.2 (Length of AGTG for a single gene has finite moments)**.** *Let $\mathcal{A}_n$ be the AGTG for a single gene from Definition 4.1 and let $L(\mathcal{A}_n)$ be its length. Then, $\mathbf{E}[L(\mathcal{A}_n)^k] < \infty$ for all $k = 1, 2, ...$*

Proof. The number of lines in $\mathcal{A}_n$ is a birth-death process with birth rate $\widehat{\lambda}_i = \gamma i/2$ and death rate $\widehat{\mu}_i = \binom{i}{2} + \rho i/2$ (when there are i lines) and 0 as absorbing state. Since the length during times with i lines increases at rate i, $L(\mathcal{A}_n)/2$ is distributed as the hitting time T of 0 of a birth-death process $(Z_t)_{t\geq 0}$ with rates $\lambda_i = \gamma$ and $\mu_i = i - 1 + \rho$, $i = 1, 2, ...$ and absorbing state 0. In order to show finite moments of T, note that the process $(\check{Z}_t)_{t\geq 0}$ with $\check{Z}_t = Z_t - 1$ is bounded from above by a birth-death process with birth rates $\check{\lambda}_k = \lambda$ and death-rate $\check{\mu}_k = k$. In other words, $(\check{Z}_t)_{t\geq 0}$ is the number of customers in an $M/M/\infty$-queue. Let S denote the partial busy period of this queue (i.e. the first time when the queue is empty). Moreover, when $\check{Z}_t = 0$ we have that $Z_t = 1$ and there is a chance $\rho/(\lambda+\rho)$ that T is reached after an $\exp(\lambda+\rho)$-distributed time. From this construction, we see that $T \leq S_1 + \cdots + S_N$ where $S_k \stackrel{d}{=} S + S'$, and $S' \sim \exp(\lambda + \rho)$ independent from S, and $N \sim \text{geom}(\rho/(\lambda + \rho))$, all S_k's being independent. Hence, the assertion follows from finite moments of S (Artalejo and Lopez-Herrero, 2001). □

We now come to the desired connection between the Moran model with horizontal gene transfer and the AGTG.

Lemma 4.3 (Gene distribution of Moran model and AGTG coincide)**.** *Fix $u \in [0, 1]$ and let $\mathcal{G}_1^N, ..., \mathcal{G}_n^N$ be as in Remark 2.2. Then, for $N \to \infty$, the distribution of $\mathcal{G}_1^N(u), ..., \mathcal{G}_n^N(u)$, conditioned on $\bigcup_{i=1}^n \mathcal{G}_i^N(u) \neq \emptyset$, converges weakly to the distribution of $(G_1, ..., G_n)$ from Definition 4.1.*

Proof. Consider the graphical construction of a Moran model with horizontal gene flow from Definition 2.1, run between times $-\infty$ and 0. Let $\mathcal{G}_i^N(-t)$ be the finite measure describing the genome of individual i at time $-t$. If we consider only a single gene $u \in [0, 1]$, we can use the following procedure in order to obtain the genealogy and distribution of gene u in $\mathcal{G}_1^N, ..., \mathcal{G}_n^N$:

1. Restrict the Moran model to (i) resampling events, (ii) potential gene loss events of gene u at rate $\rho/2$ per line, (iii) potential horizontal gene transfer events of gene u at rate $\gamma/2N$ per pair of individuals.

2. Put gene gain events on all lines according to a Poisson point process with intensity $(\theta/2)du$.

Clearly, by 1. we can determine the coordinates $(i, -t)$ with the property, that $u \in \bigcup_{j=1}^n \mathcal{G}_j^N(0)$ iff $u \in \mathcal{G}_i^N(-t)$. This subgraph is a random graph, which can be constructed from time 0 backwards as follows: Starting with n lines, any pair of lines coalesces at rate 1, every line is killed at rate $\rho/2$ and every line splits in two lines (due to a horizontal gene transfer event) at rate $\gamma(N - i)/2N$ if there are currently i lines within the graph. (For the latter rate, observe that the donor of gene u might already be part of the graph.) Hence, as $N \to \infty$, this random graph converges (weakly) to the AGTG for a single gene as in Definition 4.1. In addition, for small ε, genes in $(u - \varepsilon, u + \varepsilon)$ are gained at most once on this graph. Hence, when conditioning on the event of a gene gain of gene u on the random graph (i.e. the Poisson point process has a point (x, u)), by well-known

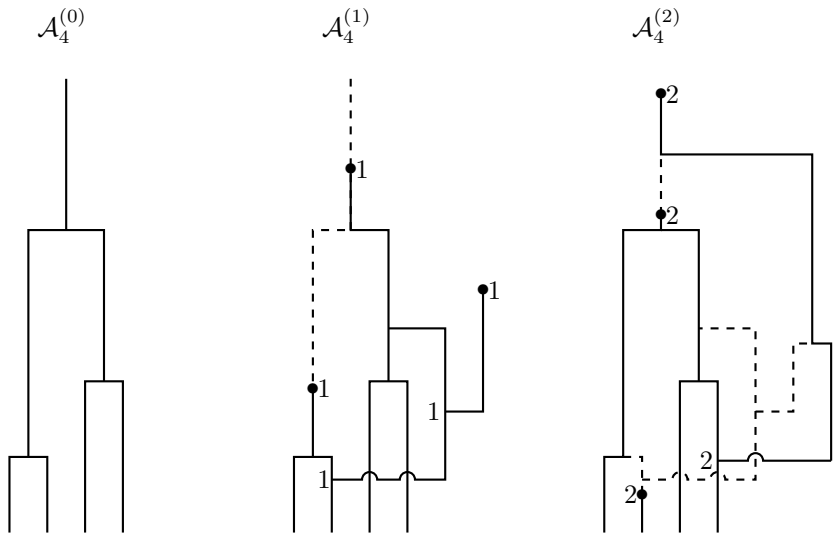

Figure 3: In the construction of the AGTG, start with the clonal genealogy of the sample (here of size 4), i.e. with $\mathcal{A}_4^{(0)}$. Then, in order to obtain the genealogy of the first gene, construct $\mathcal{A}_4^{(1)}$ by additional splitting events (at rate $\gamma/2$), loss events (at rate $\rho/2$), both marked by 1, and coalescence events. Iteratively, construct $\mathcal{A}_4^{(n+1)}$ by keeping all lines in $\cup_{i=0}^{n}\mathcal{A}_4^{(i)}$ and adding splitting, loss and coalescence events. In the three figures, the solid lines are the ones where – potentially – the corresponding gene can be gained. This means that gene 2 is present only if the second gain event at time T_2 occurs at a time which is smaller than the sum of all (vertical) lines in $\mathcal{A}_4^{(2)}$, i.e. smaller than $L(\mathcal{A}_4^{(2)})$. If it occurs, it is put uniformly on the solid lines.

properties of Poisson processes, this event is uniformly distributed on the graph. In other words, the distribution is the same as that of E in Definition 4.1. □

While the construction of the genealogy of a single gene was straight-forward, considering all genealogies of all genes seems to be harder. The reason is that there can be infinitely many genes, and each of these genes comes with its own events of gene gain, loss and horizontal gene flow. Even worse, we can decide on the presence or absence of a gene only if we know if there was a gene gain event somewhere along the genealogy, which means that we have to follow all (uncountably many) potential genes back in time.

We will resolve such difficulties by constructing (countably) many potential genealogies and model gene gain events along them. The result is the ancestral gene transfer graph (AGTG) for infinitely many genes. An illustration is found in Figure 3.

Definition 4.4 (The AGTG for infinitely many genes)**.** *Consider a sequence $\mathcal{A}_n^{(0)}, \mathcal{A}_n^{(1)}, \mathcal{A}_n^{(2)}, ...$ of coupled random graphs which arise as follows:*

- *$\mathcal{A}_n^{(0)}$ is distributed according to Kingman's coalescent, i.e. starting with n lines, each (unordered) pair of lines coalesces at rate 1 (and the graph is*

stopped upon reaching 1 line).

Given $\mathcal{A}_n^{(0)}$, the random graph $\mathcal{A}_n^{(1)}$ gives all potential ancestors of gene 1 and is constructed such that: Starting with the same n lines as in $\mathcal{A}_n^{(0)}$,

- *each line splits in a continuing and an incoming line at rate $\gamma/2$ (meaning that the gene was horizontally transferred from the incoming line). If the line was part of $\mathcal{A}_n^{(0)}$, the continuing line runs along $\mathcal{A}_n^{(0)}$ as well. The resulting splitting event is marked with "1"*
- *each line is terminated by a loss event, also marked with "1", at rate $\rho/2$ (indicating that gene 1 was lost).*
- *each (unordered) pair of lines in $(\mathcal{A}_n^{(1)} \setminus \mathcal{A}_n^{(0)})^2$ and in $\mathcal{A}_n^{(0)} \times (\mathcal{A}_n^{(1)} \setminus \mathcal{A}_n^{(0)})$ coalesces at rate 1 (and $\mathcal{A}_n^{(1)}$ is stopped upon reaching 1 line).*

Given $\mathcal{A}_n^{(0)}, ..., \mathcal{A}_n^{(k)}$, the random graph $\mathcal{A}_n^{(k+1)}$ gives all potential ancestors of gene $k+1$ and is constructed such that: Starting with the same n lines as in $\mathcal{A}_n^{(0)}$,

- *each line splits in a continuing and an incoming line at rate $\gamma/2$ (meaning that the gene was horizontally transferred from the incoming line). If the line was part of $\bigcup_{j=0}^k \mathcal{A}_n^{(j)}$, the continuing line runs along $\bigcup_{j=0}^k \mathcal{A}_n^{(j)}$ as well. The resulting splitting event is marked with "$k+1$".*
- *each line is terminated by a loss event, also marked with "$k+1$", at rate $\rho/2$ (indicating that gene 1 was lost).*
- *each (unordered) pair of lines in $(\mathcal{A}_n^{(k+1)} \setminus \bigcup_{j=0}^k \mathcal{A}_n^{(j)})^2$ and in $\bigcup_{j=0}^k \mathcal{A}_n^{(j)} \times (\mathcal{A}_n^{(k+1)} \setminus \bigcup_{j=0}^k \mathcal{A}_n^{(j)})$ coalesces at rate 1 (and the graph $\mathcal{A}_n^{(k+1)}$ is stopped upon reaching 1 line).*

In order to model gene gain events, consider the events $(T_m, U_m)_{m=1,2,...}$ of a Poisson point process on $[0,\infty) \times [0,1]$ with intensity measure $\frac{1}{2}\theta\, dt\, du$ (ordered by their first coordinate). For all k,

- *let $L(\mathcal{A}_n^{(k)})$ be the length of $\mathcal{A}_n^{(k)}$. If $T_k \leq L(\mathcal{A}_n^{(k)})$, pick a point E_k uniformly at random according to the length measure on $\mathcal{A}_n^{(k)}$. (This point determines the time and line hen the gene under consideration was gained.)*

Finally, for every $i = 1, ..., n$, let $U_k \in \mathcal{G}_i$ if there is a direct (i.e. increasing in time) path from i to E_k. Then, $(\mathcal{G}_1, ..., \mathcal{G}_n)$ is denoted the Gene distribution read off from the AGTG in the infinitely many genes model.

Remark 4.5 (Alternative way of distribution gain events on the AGTG)**.** In the last step of constructing the AGTG, we used the condition $T_k \leq L(\mathcal{A}_n^{(k)})$ in order to distribute a uniformly chosen point E_k on $\mathcal{A}_n^{(k)}$. In distribution, the same result is achieved as follows: If $T_k \leq L(\mathcal{A}_n^{(k)})$, choose a way of running through $\mathcal{A}_n^{(k)}$ along all paths at constant speed. Then, the gene gain event is placed after running length T_k.

The following Lemma is the key element in the proofs given in Section 5.

Lemma 4.6 (Gene distribution from Moran model and AGTG coincide)**.** *Fix $n \in \mathbb{N}$, let $(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)})$ be as in Definition 2.1 and Remark 2.2, and $(\mathcal{G}_1, ..., \mathcal{G}_n)$ as in Definition 4.4. Then,*

$$(\mathcal{G}_1^N, ..., \mathcal{G}_n^{(N)}) \xRightarrow{N\to\infty} (\mathcal{G}_1, ..., \mathcal{G}_n) \tag{4.1}$$

as well as

$$(\mathcal{G}_1^N \otimes \cdots \otimes \mathcal{G}_n^N) \xRightarrow{N\to\infty} (\mathcal{G}_i \otimes \cdots \otimes \mathcal{G}_n) \tag{4.2}$$

Remark 4.7 (Interpretation of (4.1) and (4.2) and a convergence criterion)**.**

1. Note that the space of finite (counting) measures on $[0, 1]$ is equipped with the topology of weak (or vague) convergence. In addition, we will interpret a vector $(\xi_1, ..., \xi_n)$ of counting measures on $[0, 1]$ as a counting measure on $\{1, ..., n\} \times [0, 1]$. Henceforth, we write $\mathcal{G}^N(\cup_{i=1}^n \{i\} \times A_i) = \prod_{i=1}^n \mathcal{G}_i^N(A_i)$ for $A_1, ..., A_n \in \mathcal{B}([0, 1])$ such that (4.1) is the same as

   $$(\langle \mathcal{G}_1^N, f_1\rangle, ..., \langle \mathcal{G}_n^{(N)}, f_n\rangle) \xRightarrow{N\to\infty} (\langle \mathcal{G}_1, f_1\rangle, ..., \langle \mathcal{G}_n, f_n, \rangle)$$

   for all $f_1, ..., f_n \in \mathcal{C}([0, 1])$.

   Since $1 \in \mathcal{C}([0, 1])$, (4.1) also implies the convergence of total masses of $\mathcal{G}_i^N$. In addition, (4.2) is stronger because total masses of products, $\langle \mathcal{G}_1^N, 1\rangle \cdots \langle \mathcal{G}_n^N, 1\rangle$ converge as well.

2. In our proof, we use the following convergence criterion from Kallenberg (2002), Proposition 16.17, here adapted for random measures on a compact space:

   > Let $\xi, \xi^1, \xi^2, ...$ be random counting measures on a compact metric space I, where ξ is simple. Then, $\xi_n \xRightarrow{N\to\infty} \xi$, if (i) $\mathbf{P}(\xi_n(A) = 0) \xrightarrow{n\to\infty} \mathbf{P}(\xi(A) = 0)$ for all open $A \subseteq I$ and (ii) $\limsup_{n\to\infty} \mathbf{E}[\xi_n(A)] \leq \mathbf{E}[\xi(A)] < \infty$ for all compact $A \subseteq I$.

Proof of Lemma 4.6. We proceed in five steps. In Step 1, we define another set of models for a population of size N with horizontal gene transfer, indexed by K, in which $I = [0, 1]$ is separated into K classes of genes, $\Delta_i^K = [(i - 1)/K; i/K), i = 1, ..., K$. For the resulting genomes, denoted $(\mathcal{G}_1^{N,K}, ..., \mathcal{G}_n^{N,K})$, we show in Step 2 that the genealogies of $(\mathcal{G}_1^{N,K}, ..., \mathcal{G}_n^{N,K})$ are given by an AGTG with $K + 1$ coupled random graphs. The construction of these random graphs can be re-ordered such that the limit $K \to \infty$ can be taken easily; see Step 3. In Step 4, we let $N \to \infty$ and show the convergence of the coupled random graphs to $(\mathcal{A}^{(0)}, \mathcal{A}^{(1)}, ...)$, implying the convergence to $(\mathcal{G}_1, ..., \mathcal{G}_n)$. In the last step we show convergence of second moments.

Step 1: Definition of $\mathcal{G}_i^{N,K}$: Fix $K \in \mathbb{N}$ and set $\Delta_i^K := [(i - 1)/K; i/K), i = 1, ..., K$. We define another Moran model (called Moran$_\Delta$-model) with horizontal gene transfer. Briefly, in this model, all genes $u \in \Delta_i^K$ follow the same gene loss and gene transfer events. Precisely, in addition to resampling events (at rate 1 for every ordered pair of individuals), the following events occur:

1. *Gene loss*: For all $k = 1, ..., K$, a gene loss event occurs at rate $\gamma/2$ per individual. Upon such an event in individual i, we have $\mathcal{G}_i^{N,K}(t) = \mathcal{G}_i^{N,K}(t-) \setminus \Delta_k^K$, i.e. all genes $u \in \Delta_k^K$ are lost from $\mathcal{G}_i^{N,K}(t)$.

2. *Gene gain*: For every individual i, at rate $\theta/2$, choose U uniform in $[0, 1]$. If $U \in \Delta_k^K$, set $\mathcal{G}_i^{N,K}(t) = (\mathcal{G}_i^{N,K}(t-) \setminus \Delta_k^K) \cup \{U\}$, i.e. U is the only gene in $\mathcal{G}_i^{N,K}(t) \cap \Delta_k^K$.

3. *Horizontal gene transfer*: For every (ordered) pair of individuals (i, j) and $k = 1, ..., K$, a horizontal gene transfer event occurs at rate $\gamma/2N$. For such an event, set $\mathcal{G}_j^{N,K}(t) = \mathcal{G}_j^{N,K}(t-) \cup (\mathcal{G}_i^{N,K}(t-) \cap \Delta_k^K)$, i.e. individual i is the donor of all genes $u \in \mathcal{G}_i^{N,K}(t-) \cap \Delta_k^K$ to the acceptor j.

Again, $(\mathcal{G}_1^{N,K}(t), ..., \mathcal{G}_N^{N,K}(t))_{t \geq 0}$ is a Harris recurrent Markov chain. We start it at time $-\infty$ and thus obtain the equilibrium measures $(\mathcal{G}_1^{N,K} := \mathcal{G}_1^{N,K}(0), ..., \mathcal{G}_N^{N,K} := \mathcal{G}_N^{N,K}(0))$ by time 0.

Step 2: $(\mathcal{G}_1^{N,K}, ..., \mathcal{G}_n^{N,K})$ can be constructed using $K + 1$ random graphs: Recall the construction of the AGTG for a single gene from Definition 4.1. We extend this construction in order to obtain the distribution of $(\mathcal{G}_1^{N,K}, ..., \mathcal{G}_n^{N,K})$. Since K is finite, we can proceed by a two-step procedure similar to the proof of Lemma 4.3 in the Moran$_\Delta$ model. Here, we first generate resampling, gene loss and transfer events and only afterwords introduce gene gain events. So, first consider a Moran model with (i) resampling events, (ii) potential gene loss events for genes in Δ_k^K with rate $\rho/2$ along all lines, where a transition from $\mathcal{G}_i^{N,K}(t)$ to $\mathcal{G}_i^{N,K}(t) \setminus \Delta_k^K$ occurs, $k = 1, ..., K$ and (iii) potential gene transfer events of genes in Δ_k^K with rate $\gamma/2N$ per pair of individuals, as in Moran$_\Delta$, $k = 1, ..., K$. Next, introduce gene gain events for all lines and all $\Delta_k, k = 1, ..., K$ at rate $\theta/2K$, where each new gene is assigned a uniformly distributed random variable on Δ_k^K.

Equivalently, as for the AGTG for a single gene, we can start from time 0 backwards and construct $K + 1$ random graphs such that graph k describes the possible ancestry of genes in $\Delta_k^K, k = 1, ..., K$. Precisely, start with graph 0, which is a coalescent started with n individuals (without gene loss and horizontal gene transfer events). In graph 1, add gene loss events, valid for all $u \in \Delta_1^K$ and gene transfer events, which lead to splits of lines in the graph at rate $\gamma(N - m)/2N$, if it currently has m lines. In addition, at rate $\gamma m/2N$, the split of a line leads to ancestry to a line which is already within the graph. Iteratively, in graph k, additional loss and split events, valid for genes $u \in \Delta_k^K$, occur. Again, a split might origin from a line which was already present in graph $0, ..., k - 1$, and otherwise gives a new line. These graphs are denoted $\mathcal{A}_{n,N,K}^{(0)}, ..., \mathcal{A}_{n,N,K}^{(K)}$.

After having constructed all $K + 1$ random graphs, graphs $1, ..., K$ are hit by gene gain events, each with rate $\theta/2K$. As above, each new gene in graph k is assigned a uniformly distributed random variable on Δ_k^K. In each Δ_k^K, keep only the gene which is closest to time 0, since in Moran$_\Delta$, a new gene in Δ_k^K overwrites present ones. By this procedure, we can read off $(\mathcal{G}_1^{(N,K)}, ..., \mathcal{G}_n^{(N,K)})$ from the random graphs, which are marked by gene gain events. Note that $\mathcal{G}_i^{N,K}(\Delta_k^K) \leq 1$ and that graphs $1, ..., K$ are exchangeable by construction.

Step 3: $(\mathcal{G}_1^{N,K},...,\mathcal{G}_n^{N,K}) \overset{K\to\infty}{\Longrightarrow} (\mathcal{G}_1^N,...,\mathcal{G}_n^N)$ *and the limit can be constructed using infinitely many random graphs*: In the construction of the last step, we reverse the order of generating gene gain events and the random graphs. First, let $(T_m,U_m)_{m=1,2,...}$ be the points in a Poisson point process $\mathcal{T}$ on $[0,\infty)\times[0,1]$ with intensity $\frac{\theta}{2}dtdu$, ordered by their first coordinates. Instead of constructing the random graphs $0,...,K+1$ in the order of the intervals $\Delta_1^N,...,\Delta_K^N$ in $[0,1]$, we can as well construct the random graphs in the order of appearance of gene gain events. Formally, let $K_m := k$ if $U_m \in \Delta_k^K$, i.e. K_m gives the number of the interval Δ_k^K in which the mth mutation $(T_m,U_m)_{m=1,2,...}$ appears. Then, let $i_1 := 1$ and $i_{r+1} := \inf\{m > i_r : K_m \notin \{K_1,...,K_{m-1}\}$ for $r < K$. This means that $K_{i_1},...,K_{i_K}$ is the number of intervals in the order of the appearance of the first gene gain within each interval. Most importantly, $(\mathcal{A}_{n,N,K}^{(0)},\mathcal{A}_{n,N,K}^{(1)}...,\mathcal{A}_{n,N,K}^{(K)}) \overset{d}{=} (\mathcal{A}_{n,N,K}^{(0)},\mathcal{A}_{n,N,K}^{(K_{i_1})},...,\mathcal{A}_{n,N,K}^{(K_{i_K})})$ since the Poisson point process $\mathcal{T}$ is independent of $(\mathcal{A}_{n,N,K}^{(1)},...,\mathcal{A}_{n,N,K}^{(K)})$ and the random graphs $(\mathcal{A}_{n,N,K}^{(1)},...,\mathcal{A}_{n,N,K}^{(K)})$ are exchangeable.

Now, consider gene gain events on $\mathcal{A}_{n,N,K}^{(K_1)}$. By construction, the first gene gain event at time T_1 falls into $\Delta_{K_1}^K$. Hence, this graph is hit after an exponentially distributed time with rate $\theta/2$. (Note the difference to the rate $\theta/2K$ from the last step.) In order to model this, take T_1 (the time of the first gene gain in $\mathcal{T}$), determine a set of paths how to move through $\mathcal{A}_{n,N,K}^{(K_1)}$ and place a gene gain event after time T_1. In case the length of $\mathcal{A}_{n,N,K}^{(K_1)}$ is smaller than T_1, do nothing. Continuing, by construction, $\mathcal{A}_{n,N,K}^{(K_{i_2})}$ (recall $K_{i_2} = K_2$ if $K_2 \neq K_1$) is hit by a gene gain event, this event occurs by time T_2 of $\mathcal{T}$. Again, determine a set of paths how to move through $\mathcal{A}_{n,N,K}^{(K_2)}$ and place a gene gain event after time T_2, if possible. Continue until T_{i_K} and $\mathcal{A}_{n,N,K}^{(K_K)}$.

In this construction, we can now let $K\to\infty$, which means that we construct infinitely many random graphs, $\mathcal{A}_{n,N}^{(0)},\mathcal{A}_{n,N}^{(1)},...$ such that the first $K+1$ are distributed according to $(\mathcal{A}_{n,N,K}^{(0)},\mathcal{A}_{n,N,K}^{(K_{i_1})},...,\mathcal{A}_{n,N,K}^{(K_{i_K})})$. On these infinitely many random graphs, we use all points (T_m,U_m) in order to construct $\widetilde{\mathcal{G}}^N$.

We now show that $\mathcal{G}^{N,K} \overset{K\to\infty}{\Longrightarrow} \widetilde{\mathcal{G}}^N$ (where the latter is constructed from the infinitely many random graphs) as well as $\mathcal{G}^{N,K} \overset{K\to\infty}{\Longrightarrow} \mathcal{G}^N$ (the latter being the set of genomes from the Moran model), implying that $\mathcal{G}^N \overset{d}{=} \widetilde{\mathcal{G}}^N$, i.e. the genomes $\mathcal{G}^N$ can be constructed from infinitely many random graphs by using all points in $\mathcal{T}$. We use the criterion from Remark 4.7. For the convergence to $\widetilde{\mathcal{G}}^N$, note that both, $\mathcal{G}^{N,K}$ and $\widetilde{\mathcal{G}}^N$ can be constructed on a joint probability space, using the same (infinitely many) random graphs. The difference in construction is that for $\widetilde{\mathcal{G}}^N$, all points in $\mathcal{T}$ are used, while in $\mathcal{G}^{N,K}$ only the first points within each Δ_i^K are used. Moreover, as long as at most one gene gain event hits $\mathcal{A}_{n,N,K}^{(i)}$ the random measures $\mathcal{G}^{N,K}$ and $\widetilde{\mathcal{G}}^N$ agree on Δ_i^K. Hence, we write, for

any Borel set $A \subseteq \{1,...,n\} \times [0,1]$ and $k = 0, 1, 2, ...$

$$
\begin{aligned}
&|\mathbf{P}(\widetilde{\mathcal{G}}^N(A) = k) - \mathbf{P}(\mathcal{G}^{N,K}(A) = k)| \\
&\qquad \leq \mathbf{P}\Big(\bigcup_{i=1}^{K} \mathcal{A}_{N,K}^{(i)} \text{ hit by 2 gene gain events} \Big) \\
&\qquad \leq K \cdot \mathbf{P}(\mathcal{A}_{N,K}^{(1)} \text{ hit by 2 gene gain events}) \\
&\qquad \leq K \cdot \mathbf{E}[1 - \exp(-\theta L(\mathcal{A}_N^{(1)})/2K)(1 + \theta L(\mathcal{A}_N^{(1)})/2K)] \\
&\qquad \leq \frac{K\theta^2}{4K^2} \mathbf{E}[(L(\mathcal{A}_N^{(1)}))^2] \xrightarrow{K\to\infty} 0,
\end{aligned}
\tag{4.3}
$$

implying (i) of Remark 4.7.2. Now, let $L(\mathcal{A}_{1,N,K}^{(1)})$ and $L(\mathcal{A}_{1,N}^{(1)})$ be the lengths of the random graphs $\mathcal{A}_{1,N,K}^{(1)}$ and $\mathcal{A}_{1,N}^{(1)}$, respectively and note that $\mathcal{A}_{1,N,K}^{(1)} \stackrel{d}{=} \mathcal{A}_{1,N}^{(1)}$. By construction, we write

$$
\begin{aligned}
\mathcal{G}_1^{N,K}([0,1]) &= \sum_{m=1}^{K} 1_{L(\mathcal{A}_{1,N,K}^{(K_m)}) \geq T_m}, \\
\mathcal{G}_1^{N}([0,1]) &= \sum_{m=1}^{\infty} 1_{L(\mathcal{A}_{1,N}^{(m)}) \geq T_m}.
\end{aligned}
\tag{4.4}
$$

Then, by exchangeability, for a rate-1-exponentially distributed random variable X and $\mathbf{E}[L(\mathcal{A}_{1,N,K}^{(1)})] = \mathbf{E}[L(\mathcal{A}_{1,N}^{(1)})] \leq \mathbf{E}[L(\mathcal{A}_1)] < \infty$ by Lemma 4.2,

$$
\begin{aligned}
\mathbf{E}[\mathcal{G}^{N,K}(\{1,...,n\} \times [0,1])] &= n \cdot \mathbf{E}[\mathcal{G}_1^{N,K}([0,1])] \\
&= nK \cdot \mathbf{P}(L(\mathcal{A}_{1,N,K}^{(1)}) \geq \tfrac{2K}{\theta} X) \\
&= nK \cdot \mathbf{E}[1 - \exp(-\theta L(\mathcal{A}_{1,N,K}^{(1)})/2K)] \\
&\xrightarrow{K\to\infty} n\theta \cdot \mathbf{E}[L(\mathcal{A}_{1,N}^{(1)})] = n \cdot \mathbf{E}[\mathcal{G}_1^{N}([0,1])] \\
&= \mathbf{E}[\mathcal{G}^{N}(\{1,...,n\} \times [0,1])],
\end{aligned}
\tag{4.5}
$$

which gives (ii) of the convergence criterion given in Remark 4.7.2. (Note that the finiteness of the right hand side of the last equation can be seen from $\mathbf{E}[L(\mathcal{A}_{1,N}^{(1)}] < \infty$; see Lemma 4.2.) Next, we come to the convergence $\mathcal{G}^{N,K} \stackrel{K\to\infty}{\Longrightarrow} \mathcal{G}^N$. Again, we observe that both random measures can be constructed on one probability space. Here, use the $K+1$ random graphs in order to construct $\mathcal{G}^{N,K}$ first and draw them as a part of a graphical construction of the Moran$_\Delta$-model, starting at time 0. Note that in the Moran$_\Delta$-model, gene gain events for a gene $u \in \Delta_k^K$ can lead to loss of another gene $v \in \Delta_k^K$, if the line of the gene gain event carries gene v. For such genes, which are lost in the Moran$_\Delta$-model, put additional gene loss and transfer events in the (regular) Moran model. Again, we claim that $\mathcal{G}^{N,K} = \mathcal{G}^N$ if every random graph $\mathcal{A}_{n,N,K}^{(1)}, ..., \mathcal{A}_{n,N,K}^{(K)}$ is hit by at most one gene gain event. Hence, the same calculations as in (4.3) and (4.5) gives the convergence $\mathcal{G}^{N,K} \stackrel{K\to\infty}{\Longrightarrow} \mathcal{G}^N$ as well.

Step 4: $(\mathcal{G}_1^N, ..., \mathcal{G}_n^N) \overset{N\to\infty}{\Longrightarrow} (\mathcal{G}_1, ..., \mathcal{G}_n)$*, constructed from infinitely many random graphs*: By now, we have shown that $(\mathcal{G}_1^N, ..., \mathcal{G}_n^N)$ can be constructed from infinitely many random graphs $\mathcal{A}_N^{(0)}, \mathcal{A}_N^{(1)}, ...$ such that $\mathcal{A}_N^{(0)}$ is a Kingman coalescent started with n lines, $\mathcal{A}_N^{(i+1)}$ has additional coalescence events and split events at rate $\gamma(N-k)/2N$, if there are k lines in graphs $\mathcal{A}_N^{(0)}, ...\mathcal{A}_N^{(i)}$. Now, as $N \to \infty$, the splitting rate converges to $\gamma/2$. By weak convergence of the random graphs, the genomes converge as well, i.e. $(\mathcal{G}_1^N, ..., \mathcal{G}_n^N) \overset{N\to\infty}{\Longrightarrow} (\mathcal{G}_1, ..., \mathcal{G}_n)$. Precisely, we again have to check (i) and (ii) of Remark 4.7.2. For (i), first note that

$$\begin{aligned}
&\mathbf{P}(\mathcal{G}_k([0,1]) \geq C) \leq \frac{1}{C}\mathbf{E}[\mathcal{G}_k([0,1])] = \frac{\theta}{2C}\mathbf{E}[L(\mathcal{A}_n^{(1)})] \xrightarrow{C\to\infty} 0, \\
&\mathbf{P}\Big(\mathcal{A}_N^{(1)} \text{ hit by split event at rate } \gamma m/2N\Big) \\
&\qquad = 1 - \mathbf{E}[\exp(-\gamma L(\mathcal{A}_N^{(1)})/2N)] \leq \mathbf{E}[\gamma L(\mathcal{A}_N^{(1)})/2N)] \xrightarrow{N\to\infty} 0.
\end{aligned} \tag{4.6}$$

according to Lemma 4.2. So, we write for $A \subseteq \{1, ..., n\} \times [0,1]$

$$\begin{aligned}
&|\mathbf{P}(\mathcal{G}(A) = k) - \mathbf{P}(\mathcal{G}^N(A) = k)| \\
&\quad \leq \mathbf{P}(\mathcal{G}(A) \geq C) + \mathbf{P}\Big(\bigcup_{i=1}^{C} \mathcal{A}_N^{(i)} \text{ hit by split event at rate } \gamma m/2N\Big) \\
&\quad \leq \mathbf{P}(\mathcal{G}(A) \geq C) + C \cdot \mathbf{P}(\mathcal{A}_N^{(1)} \text{ hit by split event at rate } \gamma m/2N\Big) \\
&\quad \xrightarrow{N\to\infty} \mathbf{P}(\mathcal{G}(A) \geq C) \xrightarrow{C\to\infty} 0
\end{aligned} \tag{4.7}$$

by (4.6) implying (i) of Remark 4.7.2. For (ii) (again noting that the same calculation holds for arbitrary compact $A \subseteq \{1, ..., n\} \times [0,1]$), we have, since $(L(\mathcal{A}_N^1))_{N=1,2,...}$ is uniformly integrable, by standard arguments (see e.g. Billingsley (1999), Theorem 3.5),

$$\begin{aligned}
&\mathbf{E}[\mathcal{G}^N(\{1, ..., n\} \times [0,1])] = n\theta \cdot \mathbf{E}[L(\mathcal{A}_{1,N}^{(1)})] \\
&\qquad \xrightarrow{N\to\infty} n\theta \cdot \mathbf{E}[L(\mathcal{A}_1^{(1)})] = \mathbf{E}[\mathcal{G}(\{1, ..., n\} \times [0,1])] < \infty.
\end{aligned} \tag{4.8}$$

Step 5: Convergence of moments: The calculations are similar to (4.7) and (4.8). We only have to deal with finiteness of moments in order to show $\mathcal{G}_1^N \otimes \cdots \otimes \mathcal{G}_n^N \overset{N\to\infty}{\Longrightarrow} \mathcal{G}_1 \otimes \cdots \otimes \mathcal{G}_j$. Here, (i) of Remark 4.7.2 is implied by (4.7). For (ii), we know that $(L(\mathcal{A}_N^{(1)}) \cdots L(\mathcal{A}_N^{(n)}))_{N=1,2,...}$ is uniformly integrable by Lemma 4.2 (since $L(\mathcal{A}_N^{(1)}) \cdots L(\mathcal{A}_N^{(n)}) \leq L(\mathcal{A}_N^{(1)})^n + \cdots + L(\mathcal{A}_N^{(n)})^n$ and the latter is uniformly integrable by Lemma 4.2) and $L(\mathcal{A}_N^{(1)}) \cdots L(\mathcal{A}_N^{(n)}) \overset{N\to\infty}{\Longrightarrow} L(\mathcal{A}_1) \cdots L(\mathcal{A}_n)$. Hence,

$$\begin{aligned}
&\mathbf{E}[\mathcal{G}_1^N \otimes \cdots \otimes \mathcal{G}_n^N((\{1, ..., n\} \times [0,1]))^n] = \tfrac{\theta^n}{2^n} \cdot \mathbf{E}[L(\mathcal{A}_{1,N}^{(1)}) \cdots L(\mathcal{A}_{1,N}^{(n)})] \\
&\quad \xrightarrow{N\to\infty} \tfrac{\theta^n}{2^n} \cdot \mathbf{E}[L(\mathcal{A}_1) \cdots L(\mathcal{A}_n)] = \mathbf{E}[\mathcal{G}_1 \otimes \cdots \otimes \mathcal{G}_n((\{1, ..., n\} \times [0,1])^n)] \\
&\quad < \infty.
\end{aligned} \tag{4.9}$$

□

5 Proofs of Theorems 1–3

5.1 Proof of Theorem 1

Using diffusion theory and Lemma 2.5, we obtain first moments of all of the statistics $G_1^{(n)}, \ldots, G_n^{(n)}$ in equilibrium. Moreover, the statistics as considered in Theorem 2 are linear combinations of $G_1^{(n)}, \ldots, G_n^{(n)}$; see the first proof of Theorem 2 below.

We consider the diffusion (2.1) with infinitesimal mean and variance

$$\mu(x) = -\tfrac{\rho}{2}x + \tfrac{\gamma}{2}x(1-x), \qquad \sigma^2(x) = x(1-x).$$

The Green function for the diffusion, measuring the time the diffusion, i.e. a gene, spends in frequency x until eventual loss, if the current frequency is $\delta \leq x$, is given by

$$G(\delta, x) = 2\frac{\phi(\delta)}{\sigma^2(x)\psi(x)},$$

where

$$\psi(y) := \exp\left(-2\int_0^y \frac{\mu(z)}{\sigma^2(z)}dz\right) = (1-y)^{1-\rho}e^{-\gamma y},$$
$$\phi(x) := \int_0^x \psi(y)dy.$$

Following Durrett (2008), we introduce new genes in frequency $\delta \ll 1$ at rate $\frac{\theta}{2}\frac{1}{\phi(\delta)}$ in a consistent way. That is, the gene raises in frequency to $\varepsilon > \delta$ with probability $\frac{\phi(\delta)}{\phi(\epsilon)}$. Hence the number of genes in frequency x is Poisson with mean

$$\frac{\theta}{2}\frac{1}{\phi(\delta)}G(\delta, x) = \theta\frac{e^{\gamma x}}{x(1-x)^{1-\rho}}.$$

The gene frequency spectrum is now given by

$$\begin{aligned}
\mathbb{E}[G_k^{(n)}] &= \binom{n}{k}\int_0^1 \theta\frac{e^{\gamma x}}{x(1-x)^{1-\rho}}x^k(1-x)^{n-k}dx \\
&= \binom{n}{k}\theta\int_0^1 e^{\gamma x}x^{k-1}(1-x)^{n-k-1+\rho}dx \\
&= \theta\binom{n}{k}(k-1)!\frac{\Gamma(n-k+\rho)}{\Gamma(n+\rho)}{}_1F_1(k; n+\rho; \gamma) \\
&= \frac{\theta}{k}\frac{(n)_{\underline{k}}}{(n-1+\rho)_{\underline{k}}}\Big(1 + \sum_{m=1}^{\infty}\frac{(k)_{\bar{m}}\gamma^m}{(n+\rho)_{\bar{m}}m!}\Big)
\end{aligned}$$

where ${}_1F_1(k; n+\rho; \gamma) = 1 + \sum_{m=1}^{\infty}\frac{(k)_{\bar{m}}\gamma^m}{(n+\rho)_{\bar{m}}m!}$ is a hypergeometric function.

5.2 Proof of Theorem 2

We give two proofs, one using diffusion theory and Theorem 1, one using the AGTG from Section 4.

Proof of Theorem 2 using Theorem 1. Given the expected gene frequency spectrum from Theorem 1, it is now easy to compute first moments of A, D and G by using, in the infinite population limit,

$$\begin{aligned} A^{(1)} &\stackrel{d}{=} G_1^{(1)}, \qquad D^{(2)} \stackrel{d}{=} \tfrac{1}{2} G_1^{(2)}, \\ G^{(n)} &= |\mathcal{G}_1| + |\mathcal{G}_2 \setminus \mathcal{G}_1| + \cdots + \Big|\mathcal{G}_n \setminus \bigcup_{i=1}^{n-1} \mathcal{G}_i\Big| \end{aligned} \tag{5.1}$$

such that

$$\begin{aligned} \mathbb{E}[A^{(n)}] = \mathbb{E}[A^{(1)}] = \mathbb{E}[G_1^{(1)}] &= \frac{\theta}{\rho}\left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(1+\rho)_{\bar{m}}}\right), \\ \mathbb{E}[D^{(n)}] = \mathbb{E}[D^{(2)}] = \tfrac{1}{2}\mathbb{E}[G_1^{(2)}] &= \frac{\theta}{1+\rho}\left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(2+\rho)_{\bar{m}}}\right), \\ \mathbb{E}[G] = \sum_{k=1}^{n} \tfrac{1}{k}\mathbb{E}[G_1^{(k)}] &= \sum_{k=1}^{n} \frac{\theta}{k} \frac{k}{k-1+\rho} \sum_{m=0}^{\infty} \frac{\gamma^m}{(k+\rho)_{\bar{m}}} \\ &= \theta \sum_{m=0}^{\infty} \gamma^m \sum_{k=0}^{n-1} \frac{1}{(k+\rho)_{\overline{m+1}}} \\ &= \theta \sum_{k=0}^{n-1} \frac{1}{k+\rho} + \theta \sum_{m=1}^{\infty} \frac{\gamma^m}{m} \sum_{k=0}^{n-1} \Big(\frac{1}{(k+\rho)_{\bar{m}}} - \frac{1}{(k+1+\rho)_{\bar{m}}}\Big) \\ &= \theta \sum_{k=0}^{n-1} \frac{1}{k+\rho} + \theta \sum_{m=1}^{\infty} \frac{\gamma^m}{m} \Big(\frac{1}{(\rho)_{\bar{m}}} - \frac{1}{(n+\rho)_{\bar{m}}}\Big). \end{aligned}$$

□

Proof of Theorem 2 using the AGTG. First we note that $A^{(1)} = G^{(1)}$ and $D^{(2)} = \frac{1}{2}(|(\mathcal{G}_1^{(1)} \cup \mathcal{G}_1^{(2)}) \setminus \mathcal{G}^{(1)}| + |(\mathcal{G}_1^{(1)} \cup \mathcal{G}_1^{(2)}) \setminus \mathcal{G}^{(2)}|)$ such that

$$\begin{aligned} \mathbb{E}[A^{(n)}] &= \mathbb{E}[A^{(1)}] = \mathbb{E}[G^{(1)}], \\ \mathbb{E}[D^{(n)}] &= \mathbb{E}[D^{(2)}] = \mathbb{E}[G^{(2)}] - \mathbb{E}[G^{(1)}], \end{aligned}$$

and it suffices to compute $\mathbb{E}[G^{(n)}]$ in the proof. We will abuse notation and write dx and dy for *infinitely small* portions of the genome. In order to compute $\mathbb{E}[G^{(n)}]$, the idea is to write $\mathcal{G}^{(n)} := (\sum_{i=1}^{n} \mathcal{G}_i) \wedge 1)$ and

$$\mathbb{E}[G^{(n)}] = \mathbb{E}\Big[\int_0^1 \mathcal{G}^n(dx)\Big] = \int_0^1 \mathbb{E}[\mathcal{G}^n(dx)] = \int_0^1 \frac{\theta}{2}\mathbb{E}[L(\mathcal{A}^n)]dx = \frac{\theta}{2}\mathbb{E}[L(\mathcal{A}^n)], \tag{5.2}$$

such that we have to compute the expected length of $\mathcal{A}^n$, the AGTG for a single gene, which we denote by $L(\mathcal{A}^n)$. Recall from the proof of Lemma 4.2 that $L(\mathcal{A}^n)/2$ has the same distribution as the hitting time T of 0 of a birth-death process $(Z_t)_{t\geq 0}$ with birth rate $\lambda_i = \gamma$ and death rate $\mu_i = i + \rho - 1$. Now, it is well known (see e.g. Karlin and Taylor, 1975) that

$$\mathbb{E}[L(\mathcal{A}^n)] = \mathbb{E}[T|Z_0 = n] = \sum_{i=1}^{\infty} p_i + \sum_{k=1}^{n-1} \left(\prod_{r=1}^{k} \frac{\mu_r}{\lambda_r}\right) \sum_{m=k+1}^{\infty} p_m \tag{5.3}$$

where

$$p_i = \frac{\lambda_1 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} = \frac{\gamma^{i-1}}{\rho(1+\rho)\cdots(i-1+\rho)} = \frac{\gamma^{i-1}}{(\rho)_{\bar{i}}}. \tag{5.4}$$

Combining (5.2) and (5.3) yields

$$\begin{aligned}
\frac{1}{\theta}\mathbb{E}[G^{(n)}] &= \frac{1}{2}\mathbb{E}[L(\mathcal{A}^n)] \\
&= \sum_{i=1}^{\infty} \frac{\gamma^{i-1}}{(\rho)_{\bar{i}}} + \sum_{k=1}^{n-1} \frac{(\rho)_{\bar{k}}}{\gamma^k} \sum_{m=k+1}^{\infty} \frac{\gamma^{m-1}}{(\rho)_{\bar{m}}} \\
&= \sum_{k=1}^{n-1} \sum_{m=k+1}^{\infty} \frac{\gamma^{m-1-k}}{(\rho+k)_{\overline{m-k}}} + \sum_{i=1}^{\infty} \frac{\gamma^{i-1}}{(\rho)_{\bar{i}}} \\
&= \sum_{m=1}^{\infty} \sum_{k=0}^{n-1} \frac{\gamma^{m-1}}{(\rho+k)_{\bar{m}}} \\
&= \sum_{k=0}^{n-1} \frac{1}{k+\rho} + \sum_{m=1}^{\infty} \gamma^m \sum_{k=0}^{n-1} \frac{1}{(\rho+k)_{\overline{m+1}}} \\
&= \sum_{k=0}^{n-1} \frac{1}{k+\rho} + \sum_{m=1}^{\infty} \frac{\gamma^m}{m} \sum_{k=0}^{n-1} \left(\frac{1}{(\rho+k)_{\bar{m}}} - \frac{1}{(\rho+k+1)_{\bar{m}}} \right) \\
&= \sum_{k=0}^{n-1} \frac{1}{k+\rho} + \sum_{m=1}^{\infty} \frac{\gamma^m}{m} \left(\frac{1}{(\rho)_{\bar{m}}} - \frac{1}{(n+\rho)_{\bar{m}}} \right).
\end{aligned} \tag{5.5}$$

According to (5.1), $\mathbb{E}[A^{(n)}]$ is readily obtained and the expected number of differences is given using (5.5) by

$$\begin{aligned}
\frac{1}{\theta}\mathbb{E}[D^{(n)}] &= \frac{1}{\theta}\big(\mathbb{E}[G^{(2)}] - \mathbb{E}[G^{(1)}]\big) = \sum_{m=1}^{\infty} \frac{\gamma^{m-1}}{(\rho+1)_{\bar{m}}} \\
&= \sum_{m=0}^{\infty} \frac{\gamma^m}{(1+\rho)_{\overline{m+1}}} = \frac{1}{1+\rho}\left(1 + \sum_{m=1}^{\infty} \frac{\gamma^m}{(2+\rho)_{\bar{m}}}\right).
\end{aligned}$$

□

5.3 Proof of Theorem 3

Again, we rely on the AGTG and use approximations if the splitting rate γ is small. Since $A^{(1)} = \int_0^1 \mathcal{G}_1(dx)$, we write

$$\mathbb{V}[A^{(1)}] = \int_0^1 \mathbb{V}[\mathcal{G}_1(dx)] + \int_0^1 \int_0^1 1_{x \neq y} \mathbb{COV}[\mathcal{G}_1(dx), \mathcal{G}_1(dy)]. \tag{5.6}$$

First, given $\mathcal{A}^1$ (the AGTG for a single gene), the expected number of gene gain events on $\mathcal{A}^1$ is Poisson distributed with parameter $(\theta/2)L(\mathcal{A}^1)dx$ such that

$$\begin{aligned}\mathbb{V}[|\mathcal{G}_1(dx)|] &= \mathbb{V}\big[\mathbb{E}[|\mathcal{G}_1(dx)|\,|\mathcal{A}^1]\big] + \mathbb{E}\big[\mathbb{V}[|\mathcal{G}_1(dx)||\mathcal{A}^1]\big] \\ &= \mathbb{V}\big[\tfrac{\theta}{2}L(\mathcal{A}^1)dx\big] + \mathbb{E}\big[\mathbb{E}[|\mathcal{G}_1(dx)|\,|\mathcal{A}^1]\big] \\ &= \frac{\theta}{2}\mathbb{E}[L(\mathcal{A}^1)]dx + \mathcal{O}(dx^2) = \mathbb{E}[A^{(1)}]dx + \mathcal{O}(dx^2) \\ &= \frac{\theta}{\rho}\Big(1 + \frac{\gamma(2+\rho)+\gamma^2}{(1+\rho)(2+\rho)} + \mathcal{O}(\gamma^3)\Big)dx + \mathcal{O}(dx^2)\end{aligned} \tag{5.7}$$

Second, for $x \neq y$,

$$\begin{aligned}\mathbb{COV}[|\mathcal{G}_1(dx)|, |\mathcal{G}_1(dy)|] &= \mathbb{COV}\big[\mathbb{E}[|\mathcal{G}_1(dx)|\,|\mathcal{A}^1, \mathcal{A}^2], \mathbb{E}[|\mathcal{G}_1(dy)|\,|\mathcal{A}^1, \mathcal{A}^2]\big] \\ &\qquad + \mathbb{E}\big[\mathbb{COV}[|\mathcal{G}_1(dx)|, |\mathcal{G}_1(dy)|\,|\mathcal{A}^1, \mathcal{A}^2]\big] \\ &= \mathbb{COV}\big[\tfrac{\theta}{2}L(\mathcal{A}^1)dx, \tfrac{\theta}{2}L(\mathcal{A}^2)dy\big] \\ &= \frac{\theta^2}{4}\mathbb{COV}[L(\mathcal{A}^1), L(\mathcal{A}^2)]dx\,dy\end{aligned}$$

since $|\mathcal{G}_1(dx)|$ and $|\mathcal{G}_1(dy)|$ are independent given $\mathcal{A}^1, \mathcal{A}^2$. Now we compute $\mathbb{COV}[L(\mathcal{A}^1), L(\mathcal{A}^2)]$ up to second order in γ. For this computation, we make use of the fact that the AGTG for two genes can be defined in analogy to the AGTG for a single gene from Definition 4.1, but with two different kind of loss and transfer events. Precisely, we consider the following random graph: starting with x lines of state *only gene 1*, y lines of state *both genes* and z lines *only gene 2*, pairs of lines coalesce at rate 1. (Note that coalescence of a line of state *only gene 1* and a line of state *only state 2* gives a single line of state *both genes*.) Lines where gene 1 (gene 2) is considered are lost at rate $\rho/2$. (If a line of state *only gene 1* (*only gene 2)* is lost, it is lost completely, while if a line of state *both genes* is lost, it turns into a line of state *only gene 2* (*only gene 1*).) Finally, every line of state *only gene 1* (*only gene 2*) is split at rate $\gamma/2$ and the new line is again of state *only gene 1* (*only gene 2*). In addition, a line of state *both genes* splits at rate γ and the new line is of state *only gene 1* and *only gene 2*, both with probability 1/2. The length of the graph of lines at states *only gene 1* (*only gene 2*) and *both genes* is denoted $L_1(t)$ ($L_2(t)$) if a sample from time t of the population is considered. We write $\mathbb{E}_{xyz}[.]$ for the expected value if the process is started as above.

It is important to note that $\mathbb{E}[L(\mathcal{A}^1)L(\mathcal{A}^2)] = \mathbb{E}_{010}[L_1(t)L_2(t)]$ for any t, since the AGTG describes the population in equilibrium. In order to compute $\mathbb{E}_{010}[L_1(t)L_2(t)]$, we use a time derivative and write

$$\begin{aligned}\mathbb{E}_{010}[L_1(t+dt)L_2(t+dt)] &= (1-(\gamma+\rho)dt)\mathbb{E}_{010}[(L_1(t)+dt)(L_2(t)+dt)] \\ &\qquad + \gamma dt \cdot \mathbb{E}_{110}[(L_1(t)+dt)(L_2(t)+dt)] + \rho dt \cdot 0\end{aligned}$$

Using that the AGTG is in equilibrium and ignoring effects of order dt^2, we

obtain (for $L_i := L_i(t)$, $i = 1, 2$)

$$\begin{aligned}
(\gamma+\rho)\mathbb{E}_{010}[L_1L_2] &= \mathbb{E}_{010}[L_1+L_2] + \gamma\cdot\mathbb{E}_{110}[L_1L_2],\\
(1+\tfrac{3}{2}\rho+\tfrac{3}{2}\gamma)\mathbb{E}_{110}[L_1L_2] &= \mathbb{E}_{110}[L_1+2L_2] + \gamma\cdot\mathbb{E}_{210}[L_1L_2] + \tfrac{1}{2}\gamma\mathbb{E}_{111}[L_1L_2]\\
&\quad + \tfrac{1}{2}\rho\cdot\mathbb{E}_{101}[L_1L_2] + (1+\tfrac{1}{2}\rho)\cdot\mathbb{E}_{010}[L_1L_2],\\
(3+2\rho)\mathbb{E}_{210}[L_1L_2] &= \mathbb{E}_{210}[L_1+3L_2] + (3+\rho)\cdot\mathbb{E}_{110}[L_1L_2]\\
&\quad + \tfrac{1}{2}\rho\cdot\mathbb{E}_{201}[L_1L_2] + \mathcal{O}(\gamma),\\
(3+2\rho)\mathbb{E}_{111}[L_1L_2] &= \mathbb{E}_{111}[2L_1+2L_2] + \mathbb{E}_{020}[L_1L_2]\\
&\quad + (2+\rho)\cdot\mathbb{E}_{110}[L_1L_2] + \rho\cdot\mathbb{E}_{201}[L_1L_2] + \mathcal{O}(\gamma),\\
(1+\gamma+\rho)\mathbb{E}_{101}[L_1L_2] &= \mathbb{E}_{101}[L_1+L_2] + \mathbb{E}_{010}[L_1L_2] + \gamma\cdot\mathbb{E}_{201}[L_1L_2],\\
(3+\tfrac{3}{2}\rho)\mathbb{E}_{201}[L_1L_2] &= \mathbb{E}_{201}[L_1+2L_2] + (1+\rho)\cdot\mathbb{E}_{101}[L_1L_2]\\
&\quad + 2\cdot\mathbb{E}_{110}[L_1L_2] + \mathcal{O}(\gamma),\\
(1+2\rho)\mathbb{E}_{020}[L_1L_2] &= \mathbb{E}_{020}[2L_1+2L_2] + \mathbb{E}_{010}[L_1L_2]\\
&\quad + 2\rho\cdot\mathbb{E}_{110}[L_1L_2] + \mathcal{O}(\gamma).
\end{aligned} \tag{5.8}$$

Note that some terms $\mathcal{O}(\gamma)$ were written which will not lead to the first two leading terms in $\mathbb{E}_{010}[L_1(t)L_2(t)]$. The expectations $\mathbb{E}_j[L_i]$ for $i = 1, 2$ and $j \in \{010, 110, 101, 210, 111, 201\}$ can readily be computed using

$$\mathbb{E}_{xyz}[L_1] = \mathbb{E}[L(\mathcal{A}^{x+y})] \text{ and } \mathbb{E}_{xyz}[L_2] = \mathbb{E}[L(\mathcal{A}^{y+z})]. \tag{5.9}$$

We use from (5.5) that

$$\mathbb{E}[L(\mathcal{A}^n)] = \sum_{k=0}^{n-1}\frac{2}{k+\rho} + 2\gamma\frac{n}{\rho(n+\rho)} + \gamma^2\frac{n(n+2\rho+1)}{\rho(\rho+1)(n+\rho)(n+\rho+1)} + \mathcal{O}(\gamma^3),$$

such that

$$\begin{aligned}
\mathbb{E}[L(\mathcal{A}^1)] &= \frac{2}{\rho}\Big(1+\frac{\gamma}{1+\rho}\Big) + \gamma^2\frac{2}{\rho(\rho+1)(\rho+2)} + \mathcal{O}(\gamma^3),\\
\mathbb{E}[L(\mathcal{A}^2)] &= \frac{2}{\rho(1+\rho)}\Big(1+2\rho+2\gamma\Big) + \mathcal{O}(\gamma^2),\\
\mathbb{E}[L(\mathcal{A}^3)] &= \frac{6\rho^2+10\rho+4}{\rho(\rho+1)(\rho+2)} + \mathcal{O}(\gamma).
\end{aligned} \tag{5.10}$$

Solving (5.8) using (5.9) and (5.10) gives

$$\begin{aligned}
\mathbb{COV}[L(\mathcal{A}^1), L(\mathcal{A}^2)] &= \mathbb{E}_{010}[L_1L_2] - \mathbb{E}[L(\mathcal{A}^1)]^2\\
&= \frac{4}{\rho(1+\rho)^2(3+2\rho)(2+7\rho+6\rho^2)}\gamma^2 + \mathcal{O}(\gamma^3).
\end{aligned} \tag{5.11}$$

Combining (5.11) with (5.7) and (5.6) gives the result.

To compute the variance for the number of differences $D^{(2)}$ we will use a similar approach. Setting $\mathcal{D}_{1,2} := (\mathcal{G}_1^N - \mathcal{G}_2^N)^+ + (\mathcal{G}_2^N - \mathcal{G}_1^N)^+$ we can write $2D^{(2)} = \int_0^1 \mathcal{D}_{1,2}(dx)$ and thus

$$\mathbb{V}[2D^{(2)}] = \int_0^1 \mathbb{V}[\mathcal{D}_{1,2}(dx)] + \int_0^1\int_0^1 1_{x\neq y}\mathbb{COV}[\mathcal{D}_{1,2}(dx), \mathcal{D}_{1,2}(dy)]. \tag{5.12}$$

Let $\mathcal{A}_2^i$ be the subgraph of $\mathcal{A}_2^{(i)}$ consisting of branches leading to either individual 1 or individual 2 but not to both. If $L(\mathcal{A}_2^i)$ denotes its length, given $\mathcal{A}_2^{(i)}$, the expected number of gene gain events leading to a difference between two given individuals is Poisson distributed with parameter $(\theta/2)L(\mathcal{A}_2^i)$ such that (compare with (5.7))

$$\begin{aligned}\mathbb{V}[|\mathcal{D}_{1,2}(dx)|] &= \mathbb{V}\big[\mathbb{E}[|\mathcal{D}_{1,2}(dx)|\,|\mathcal{A}_2^{(i)}]\big] + \mathbb{E}\big[\mathbb{V}[|\mathcal{D}_{1,2}(dx)||\mathcal{A}_2^{(i)}]\big] \\ &= \mathbb{V}\big[\tfrac{\theta}{2}L(\mathcal{A}_2^i)dx\big] + \mathbb{E}\big[\mathbb{E}[|\mathcal{D}_{1,2}(dx)|\,|\mathcal{A}_2^{(i)}]\big] \\ &= \frac{\theta}{2}\mathbb{E}[L(\mathcal{A}_2^i)]dx + \mathcal{O}(dx^2) = \frac{\theta}{2}\mathbb{E}[2D^{(2)}]dx + \mathcal{O}(dx^2) \\ &= \theta\frac{2}{1+\rho} + \frac{2\gamma}{2+3\rho+\rho^2} + \mathcal{O}(\gamma^2) + \mathcal{O}(dx^2).\end{aligned} \tag{5.13}$$

In the same way as seen below equation (5.7) we obtain, for $i \neq j$

$$\mathbb{COV}[|\mathcal{D}_{1,2}(dx)|, |\mathcal{D}_{1,2}(dy)|] = \frac{\theta^2}{4}\mathbb{COV}[L(\mathcal{A}_2^i), L(\mathcal{A}_2^j)]dx\,dy.$$

As $\mathbb{E}[L(\mathcal{A}_2^i)] \cdot \mathbb{E}[L(\mathcal{A}_2^j)]$ is already known the remaining part is to compute

$$\begin{aligned}\mathbb{E}[L(\mathcal{A}_2^i)L(\mathcal{A}_2^j)] = &\frac{32}{(1+\rho)(1+2\rho)} \\ &+ \frac{32(48+314\rho+611\rho^2+464\rho^3+120\rho^4)\gamma}{(1+\rho)(2+\rho)(1+2\rho)^2(3+2\rho)(2+3\rho)(6+5\rho)} + \mathcal{O}(\gamma^2).\end{aligned} \tag{5.14}$$

For that we will split $(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$ into two parts, $T(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$ and $S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$. Recall that there are three different types of events in $(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$, namely loss, merging lines and splitting lines. The first part, $T(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$, contains solely the times $T_1, T_2, \ldots$ between these events, while the second part, $S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$ contains the remaining information from $(\mathcal{A}^1, \mathcal{A}^2)$ on which lines split, merge and get lost, i.e. it is possible to describe the structure/topology/shape of the AGTG from $S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$. Note that given $S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$, the times $T_1 = T_1(S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})), T_2 = T_2(S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})), \ldots$ are independent exponentially distributed random variables with rates measurable with respect to $S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$. In particular, the number of lines between the kth and $(k+1)$st time in $T(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$, which lead to either one or the other of the individuals, but not to both, denoted by $D_k^i = D_k^i(S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)}))$, is $S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$-measurable and

$$L(\mathcal{A}_2^i) = \sum_k D_k^i T_k \tag{5.15}$$

Let $\mathcal{S}$ be the space of all possible shapes which can be taken by $S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$ and let $\mathcal{S}_{\gamma^2} := \{s \in \mathcal{S} : \mathbb{P}(s) \notin \mathcal{O}(\gamma^2)\}$, i.e. $\mathcal{S}_{\gamma^2}$ contains all shapes which have at most one splitting event. Within $\mathcal{S}_{\gamma^2}$, there are at most 8 events before

$(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})$ has lost all lines, so we can write

$$\begin{aligned}
\mathbb{E}[L(\mathcal{A}_2^{(i)})L(\mathcal{A}_2^{(j)})] &= \mathbb{E}[\mathbb{E}[L(\mathcal{A}_2^{(i)})L(\mathcal{A}_2^{(j)})]|S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)})]] \\
&= \sum_{s\in\mathcal{S}_{\gamma^2}} \mathbb{P}(s)\cdot\mathbb{E}[L(\mathcal{A}_2^{(i)})L(\mathcal{A}_2^{(j)})]|S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)}) = s] + \mathcal{O}(\gamma^2) \\
&= \sum_{s\in\mathcal{S}_{\gamma^2}} \mathbb{P}(s)\cdot\mathbb{E}\Big[\sum_{k=1}^{8} D_k^i T_k \sum_{k=1}^{8} D_k^j T_k | S(\mathcal{A}_2^{(i)}, \mathcal{A}_2^{(j)}) = s\Big] + \mathcal{O}(\gamma^2) \\
&= \sum_{s\in\mathcal{S}_{\gamma^2}} \mathbb{P}(s)\sum_{k=1}^{8} D_k^i(s) D_k^j(s)\mathbb{E}[T_k^2(s)] \\
&\quad + \mathbb{P}(s) \sum_{1\le k,k'\le 8; k\neq k'} D_k^i(s) D_{k'}^j(s)\mathbb{E}[T_k(s)]\mathbb{E}[T_{k'}(s)] + \mathcal{O}(\gamma^2)
\end{aligned}$$

As $\mathcal{S}_{\gamma^2}$ has more than 5000 elements we used Mathematica to compute $\mathbb{P}(s)$ – see the accompanying file – the variables $D_k^i(s)$, resp. $D_k^j(s)$, and the parameters of the exponentially distributed times $T_k(s)$ for $1 \le k \le 8$ and all $s \in \mathcal{S}_{\gamma^2}$. Combining (5.13) and (5.14) gives the result as shown in (3.10).

Acknowledgments

We thank Wolfgang Hess for fruitful discussions. The DFG is acknowledged for funding via the project PP672/2-1. The SPP 1590 funded by the DFG is acknowledged for travel support.

References

Artalejo, J. and J. Lopez-Herrero (2001). Analysis of the busy persion for the $M/M/c$ queue: an algorithmic approach. *J. Appl. Prob. 38*, 209–222.

Baumdicker, F., W. R. Hess, and P. Pfaffelhuber (2010). The diversity of a distributed genome in bacterial populations. *Ann. Appl. Probab. 20*(5), 1567–1606.

Baumdicker, F., W. R. Hess, and P. Pfaffelhuber (2012). The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol. 4*(4), 443–456.

Berg, O. G. and C. G. Kurland (2002). Evolution of microbial genomes: sequence acquisition and loss. *Mol. Biol. Evol. 19*(12), 2265–2276.

Billingsley, P. (1999). *Convergence of Probability Measures. 2nd ed.* John Wiley.

Collins, R. E. and P. G. Higgs (2012). Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol. Biol. Evol. online first*, 1–15.

Dagan, T. (2011). Phylogenomic networks. *Trends Microbiol. 19*(10), 483–491.

Dagan, T. and W. Martin (2006). The tree of one percent. *Genome Biol. 7*(10), 118–118.

Dagan, T. and W. Martin (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A. 104*(3), 870–875.

de la Cruz, F. and J. Davies (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol. 8*(3), 128–133.

Didelot, X., D. Lawson, A. Darling, and D. Falush (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics 186*(4), 1435–1449.

Doolittle, W. F. (1999). Lateral genomics. *Trends Cell Biol. 9*(12), 5–8.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution* (2nd ed.). Springer.

Ehrlich, G. D., F. Z. Hu, K. Shen, P. Stoodley, and J. C. Post (2005). Bacterial plurality as a general mechanism driving persistence in chronic infections. *Clin. Orthop. Relat. Res. 437*, 20–24.

Ewens, W. J. (2004). *Mathematical PopulationGenetics. I. Theoretical introduction* (2nd ed.). Springer.

Fisher, R. (1930). The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinburgh 50*, 205–220.

Gogarten, J. P., W. F. Doolittle, and J. G. Lawrence (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol. 19*(12), 2226–2238.

Griffiths, R. and P. Marjoram (1997). An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution, IMA volumes in Mathematics and its Applications, 87. Springer Verlag, Berlin*, pp. 257–270.

Haegeman, B. and J. S. Weitz (2012). A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics 13*, 196–196.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology 23*, 183–201.

Huson, D. H. and C. Scornavacca (2011). A survey of combinatorial methods for phylogenetic networks. *Genome Biol. Evol. 3*, 23–35.

Huson, D. H. and M. Steel (2004). Phylogenetic trees based on gene content. *Bioinformatics 20*(13), 2044–2049.

Kallenberg, O. (2002). *Foundations of modern probability. 2nd ed.* Probability and Its Applications. New York, NY: Springer.

Karlin, S. and H. M. Taylor (1975). *A first course in stochastic processes.* Academic Press London.

Kimura, M. (1964). Diffusion Models in Population Genetics. *J. Appl. Probab. 1*(2), 177–192.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics 61*, 893–903.

Kingman, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab. 19A*, 27–43.

Koonin, E. V., K. S. Makarova, and L. Aravind (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol. 55*, 709–742.

Koonin, E. V., A. R. Mushegian, M. Y. Galperin, and D. R. Walker (1997). Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol. 25*(4), 619–637.

Koonin, E. V. and Y. I. Wolf (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research 36*(21), 6688–6719.

Koonin, E. V. and Y. I. Wolf (2012). Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front Cell Infect. Microbiol. 2*, 119–119.

Krone, S. and C. Neuhauser (1997). Ancestral processes with selection. *Theo. Pop. Biol. 51*, 210–237.

Kunin, V. and C. A. Ouzounis (2003). The balance of driving forces during genome evolution in prokaryotes. *Genome Research 13*(7), 1589–1594.

Lawrence, J. G. and H. Ochman (2002). Reconciling the many faces of lateral gene transfer. *Trends in Microbiology 10*(1), 1–4.

Linz, S., A. Radtke, and A. von Haeseler (2007). A likelihood framework to measure horizontal gene transfer. *Molecular Biology and Evolution 24*(6), 1312–1319.

McDaniel, L. D., E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul (2010). High frequency of horizontal gene transfer in the oceans. *Science 330*, 50.

Medini, D., C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev. 15*(6), 589–594.

Mozhayskiy, V. and I. Tagkopoulos (2012). Horizontal gene transfer dynamics and distribution of fitness effects during microbial in silico evolution. *BMC Bioinformatics 13 Suppl 10*, 1–17.

Nakhleh, L., D. Ruths, and L.-S. Wang (2005). Riata-hgt: A fast and accurate heuristic for reconstructing horizontal gene transfer. In L. Wang (Ed.), *Computing and Combinatorics*, Volume 3595 of *Lecture Notes in Computer Science*, pp. 84–93. Springer Berlin, Heidelberg.

Neuhauser, C. and S. Krone (1997). The genealogy of samples in models with selection. *Genetics 145*, 519–534.

Novozhilov, A. S., G. P. Karev, and E. V. Koonin (2005). Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol. 22*(8), 1721–1732.

Perna, N. T., G. Plunkett, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature 409*, 529–533.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial pan-genome. *Proc. Natl. Acad. Sci. U.S.A. 102*(39), 13950–13955.

Tettelin, H., D. Riley, C. Cattuto, and D. Medini (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion Microbiol. 11*(5), 472–477.

Vogan, A. A. and P. G. Higgs (2011). The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. *Biol. Direct 6*, 1–14.

Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. U.S.A. 24*, 253–259.